

The LIUM Arabic/English Statistical Machine Translation System for IWSLT 2008

Holger Schwenk Yannick Estève Sadaf Abdul-Rauf

LIUM, University of Le Mans, FRANCE

Name.Surname@lium.univ-lemans.fr



ABSTRACT

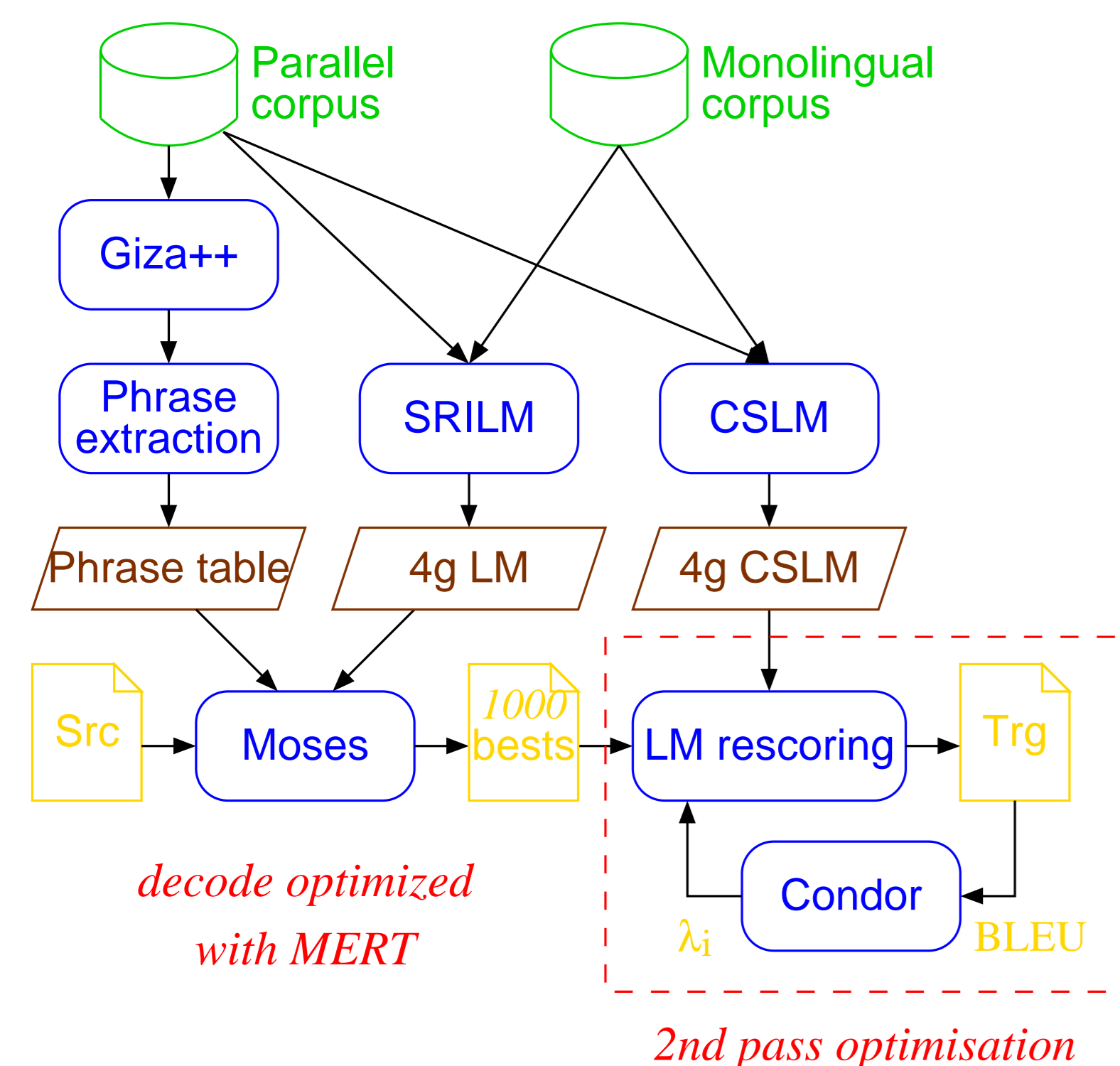
This paper describes the system developed by the LIUM laboratory for the 2008 IWSLT evaluation. We only participated in the Arabic/English BTEC task. We developed a statistical phrase-based system using the Moses toolkit and SYSTRAN's rule-based translation system to perform a morphological decomposition of the Arabic words. A continuous space language model was deployed to improve the modeling of the target language. Both approaches achieved significant improvements in the BLEU score. The system achieves a score of 49.4 on the test set of the 2008 IWSLT evaluation.

INTRODUCTION

- Only Arabic/English BTEC task (mainly text)
- Similar architecture than Ar/En NIST or Fr/En WMT system
- Only BTEC bitexts
- Small improvements using additional LM data (Gigaword)
- Two different tokenizations of the Arabic source text:
 - full word mode
 - morphological decomposition kindly provided by SYSTRAN
- No system combination

SYSTEM ARCHITECTURE

- Statistical phrase-based system using Moses and own tools
- Two pass approach:
 - Decode with Moses and generate 1000-best lists
 - Rescore n -best lists with continuous space LM
 - Maximum BLEU tuning on rescored n -best lists using public CONDOR tool
- All models are case sensitive models
- Punctuation markers are considered as normal words



Continuous space LM

- Tries to tackle the data sparseness problem [Y. Bengio, NIPS'01]
- Idea: projection of the word indices onto a continuous space
- n -gram probability estimation in this continuous space
 - ⇒ Better generalization to unseen n -grams can be expected
- Implementation using a 3-layer neural network
- Backpropagation training to learn the continuous representation of the words and the n -gram LM probabilities
- Several tricks to tackle the high complexity

Dev Data

- Dev4 and Dev5 seem to be very similar
- Dev6 is mainly close to the BTEC training corpus.
- Analysis of the Arabic source: Test08 seems to be close to Dev4/5
 - ⇒ All tuning is done on Dev05
- Results on Dev6 and Test08

Post-processing

- Translations of Test08 data contain only few punctuation marks
- This is in contrast to Dev5 and Dev6
- ⇒ negative impact on our system
- We were unable to analyze the Arabic source
- Simple post-processing to restore end-of-sentence punctuation

EXPERIMENTAL EVALUATION

Language Modeling

- English part of BTEC train and Dev1-4 (all English references)
- LDC Gigaword (3.3 billion words)
- GALE part of the 2006 NIST test set (1.1M words).
 - contains WEB blogs (tourism related ?)
 - we realized after the evaluation that this data was only distributed to participants of the NIST MT eval
- 4-gram back-off LMs with Modified Kneser-Ney smoothing
- Individual LMs are interpolated together

Corpus	train #words	LM size	Perplexity on Dev5
BTEC train	153k	3.3M	109.8
+BTEC Dev1-4	+205k	6.5M	75.0
+Gale	+1.1M	309M	71.6
+Gigaword	+3.3G	1.1G	58.4
+ CSLM	3.4G	71M	49.3

- Dev data helps a lot
- GALE data brings small improvement
- Gigaword is important although out-of-domain
- CSLM brings nice gain in perplexity

Baseline experiment with NIST Arabic/English system

Translation model	Language Model	Dev5	Dev6
NIST	NIST	21.01	33.49
NIST	BTEC+Giga	21.62	37.29
BTEC	BTEC	21.35	47.09
BTEC	BTEC+Giga	23.18	44.15

- Large News systems performs badly on BTEC tourism task
- In-domain LM improves only Dev6
- BTEC bitexts only help for Dev6
 - ⇒ The generic system achieves reasonable scores on Dev5 only

Adding more parallel data

Translation model	Language model	Dev5	Dev6	Test08
Default tokenization:				
BTEC	BTEC	21.35	47.09	43.45
	BTEC + Dev1-4	22.90	45.16	42.98
	BTEC + Dev1-4 + Giga	23.18	44.15	43.70
BTEC + Dev1-4	BTEC + Dev1-4	28.15	47.33	42.71
	BTEC + Dev1-4 + Giga	28.39	47.62	44.19
BTEC + Dev1-4 + Gale	BTEC + Dev1-4 + Giga	28.17	47.82	43.52
	larger word list	30.49	49.51	45.08
Improved tokenization:				
BTEC + Dev1-4	BTEC + Dev1-4 + Giga	31.20	52.10	48.09
	idem CSLM	32.38	52.42	47.52
BTEC + Dev1-4 + Gale	BTEC + Dev1-4 + Giga	31.63	50.76	47.16
BTEC + Dev1-6	BTEC + Dev1-6 + Giga	-	-	48.04
	idem CSLM	-	-	49.39

- The large LM with the Gigaword data has only a small impact on the BLEU scores, despite a good gain in perplexity
- Gale bitext seem to be useful

IMPROVED TOKENIZATION

- It is known that a morphological decomposition of the Arabic words can improve the word coverage and by these means the translation quality
- Particularly true for under-resourced tasks like BTEC
- Usually the Buckwalter transliterator and the MADA and TOKAN tools from Columbia University are used

Using SYSTRAN's sentence analysis

- Sentence analysis represents a large share of the computation in a rule-based system
- Apply first decomposition rules coupled with a word dictionary
- For words that are not known in the dictionary, the most likely decomposition is guessed

- In general, all possible decompositions of each word are generated and then filtered in the context of the sentence.
- This steps uses **lexical knowledge** and a **global analysis** of the sentences.

→ Integration of linguistic knowledge, but difficult to apply onto a word lattice from ASR

Result analysis:

- Substantial improvements in the BLEU score
 - Dev6: 47.62 → 52.10, Test08: 44.19 → 48.09
- Gale bitexts are not useful any more
- The morphological decomposition seems to achieve better translations than adding additional bilingual out-of domain data.

Relation to SPE:

- word based system: SMT performs the full translation task
- SPE: SMT only corrects the output of rule-based system
- SYSTRAN's tokenisation + SMT: somewhere in the continuum between both

INTERFACE WITH SPEECH RECOGNITION

- Simple 1-best coupling
 - Bad performance on ASR transcriptions

Condition	Dev5	Dev6	Test08
Text input	32.38	52.42	49.39
ASR 1-best input	28.98	43.94	38.26

CONCLUSION AND PERSPECTIVES

- Based on Moses decoder
- Two extensions achieved significant improvements:
 - morphological word decomposition based on SYSTRAN's rule-based translation system
 - n -best list rescoring with a continuous space language model
- No gain with additional bitexts
- Small improvements with additional LM data

Ongoing work

- Explore unsupervised training of translation model
- Comparison of SYSTRAN's morphological decomposition with MADA/TOKAN and other standard tools

ACKNOWLEDGMENTS

This work has been partially funded by the French Government under the project INSTAR (ANR JCJC06_143038).